Е. И. Стенина

ОСНОВЫ НАУЧНЫХ ИССЛЕДОВАНИЙ

Метод группирования данных

Екатеринбург
2020

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Уральский государственный лесотехнический университет»
(УГЛТУ)

Кафедра автоматизации и инновационных технологий

Е. И. Стенина

ОСНОВЫ НАУЧНЫХ ИССЛЕДОВАНИЙ

Метод группирования данных

Методические указания по выполнению практических,
лабораторных и исследовательских работ обучающимися
по направлению 35.03.02 «Технология лесозаготовительных
и деревоперерабатывающих производств» всех форм обучения

Екатеринбург
2020

Печатается по рекомендации методической комиссии ИЛБ ФГБОУ ВО «Уральский государственный лесотехнический университет».
Протокол № 2 от 03 октября 2019 г.

Рецензент – профессор кафедры АИТ, д-р техн. наук Гороховский А. Г.

Редактор Е. Л. Михайлова
Оператор компьютерной верстки Т. В. Упорова

Подписано в печать 18.03.20		Поз. 15
Плоская печать	Формат 60×84/16	Тираж 10 экз.
Заказ №	Печ. л. 0,7	Цена руб. коп.

Редакционно-издательский отдел УГЛТУ
Сектор оперативной полиграфии УГЛТУ

ВВЕДЕНИЕ

Любое экспериментальное исследование условно можно разделить на три этапа: подготовка эксперимента, планирование и постановка опытов, обработка результатов измерений и их анализ. Множество значений результатов экспериментов (случайных величин), полученных в продублированных опытах, представляет собой статистическую совокупность, которая иногда может содержать десятки и даже сотни наблюдений. Обычная статистическая обработка такого количества результатов становится крайне трудоемким процессом. В этом случае целесообразно применять метод группирования данных.

1. ЦЕЛЬ РАБОТЫ

Методические указания предназначены для приобретения студентами практических навыков в статистической обработке результатов эксперимента, содержащих значительное количество значений.

2. СТАТИСТИЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТОВ

Множество значений случайной величины, полученных в результате эксперимента или наблюдений над объектом исследования, представляет собой **статистическую совокупность**.

Статистическая совокупность, содержащая в себе все возможные значения случайной величины, называется **генеральной статистической** совокупностью.

Выборочной статистической совокупностью (выборкой) называется совокупность, в которой содержится только некоторая часть элементов генеральной совокупности.

Основными статистическими показателями, используемыми для характеристики любой выборочной совокупности (выборки), являются ***объем выборки, выборочное среднее и выборочная дисперсия***.

Число опытов (наблюдений) $\boldsymbol{n}$, содержащееся в выборке, называется **объемом выборки**. Чем больше объем выборки, чем меньше влияние случайных ошибок, тем большей достоверностью обладают результаты, содержащиеся в выборке.

Наилучшей оценкой истинного значения выходной величины является **выборочное среднее** $\bar{y}$.

Количественной оценкой модуля величины случайной ошибки измерения, возникающей под действием факторов, неизвестных исследователю или известных, но нерегулируемых и постоянно изменяющихся, является **выборочная дисперсия** $\mathbf{S^2}$.

При повторении опытов в одинаковых условиях обычно обнаруживается закономерность в частоте появления тех или иных результатов. Некоторые значения случайной величины появляются значительно чаще других, при этом они группируются относительно некоторого значения – центра группирования. Тогда $\boldsymbol{p_i}$ **– вероятность (относительная частота**) появления i-го значения выходной величины в выборке.

График, построенный по данным статистического ряда, называют **гистограммой.** Так как сумма всех относительных частот равна 1, то площадь всей гистограммы также равна единице. С увеличением числа опытов n значение каждой частоты становится все ближе к соответствующей вероятности p_i. Это утверждение, выражающее требование статистической устойчивости частот, является важнейшей предпосылкой применения статистических методов обработки данных.

Если с увеличением числа опытов n увеличивать и количество интервалов, то ломаная, ограничивающая гистограмму сверху, приближается к некоторой кривой, называемой **кривой распределения**, или **кривой плотности вероятности**. Она является графиком соотношения между значениями данной случайной величины и их вероятностями. В теории вероятностей это соотношение называется **статистическим распределением.** Для случайных величин, имеющих разную природу, статистические распределения могут быть различными, например распределения Пуассона, Пирсона, биномиальное и др. Среди них существует **распределение**, называемое **нормальным** (гауссовским), когда предполагается, что среди достаточно большого числа рассматриваемых случайных величин нет такой, влияние которой на их сумму существенно преобладало бы по сравнению с другими значениями, содержащимися в выборке.

3. ОБЩИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ РАБОТЫ

При обработке результатов эксперимента (выборки) методом группирования данных необходимо:

– определить наименьшее и наибольшее значения выходной величины (п. 4.1);

– весь диапазон значений случайной величины в выборке необходимо разбить на интервалы (п. 4.2);

– определить границы и середины интервалов (п. 4.3);

– подсчитать количество результатов эксперимента, попадающее в каждый интервал (п. 4.4);

– рассчитать относительную частоту попадания случайной величины в каждый интервал (п. 4.5);

– рассчитать основные статистические показатели выборки (п. 4.6);

– построить гистограмму и кривую распределения случайной величины в выборке (п. 4.7);

– проанализировать полученные результаты (п. 4.8).

Расчеты должны выполняться и **оформляться в соответствии с требованиями ЕСКД**.

4. ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

С целью демонстрации методики статистической обработки результатов эксперимента будем использовать следующий пример.

Пример

В результате эксперимента по пропитке кленовых образцов были получены следующие значения общего поглощения y_i, кг/м^3:

40,2	50,0	57,7	48,3	49,5	42,9	51,6	53,4	47,7	49,0	46,6	55,8
34,1	20,4	16,7	22,8	23,8	24,0	26,7	21,3	23,6	30,4	27,4	23,0
46,1	22,5	54,2	48,7	56,4	54,0	50,8	60,7	59,4	51,2	55,5	19,2
21,0	35,4	33,2	36,9	40,4	30,5	34,7	35,5	39,1	32,2	33,1	56,0
50,6	54,4	58,9	60,6	51,8	50,2	56,3	58,3	54,2	58,4	55,7	57,1

4.1. Определение наименьшего и наибольшего значений выходной величины

Данные значения выходной величины можно определить визуально.

Пример

$y_{\min} = 16{,}7$ кг/м^3, $y_{\max} = 60{,}7$ кг/м^3.

4.2. Расчет интервалов

4.2.1. Расчет числа интервалов

Необходимое число интервалов k определяется по формуле

$$k = 1 + 3{,}2\ lg\ n, \tag{1}$$

где n – объем выборки.

Пример

$n = 60$,
$k = 1+3{,}2\ lg60 = 6{,}69$

Полученное значение числа интервалов округляется в сторону большего числа. Таким образом, $k = 7$.

4.2.2. Расчет длины интервала

Чаще всего используют интервалы равной длины. В этом случае длина каждого интервала h рассчитывается по формуле

$$h = \frac{y_{\max} - y_{\min}}{k} . \frac{y_{max} - y_{min}}{k} \tag{2}$$

Пример

$$h = \frac{60{,}7 - 16{,}7}{7} = 6{,}29 \text{ кг/м}^3.$$

4.3. Определение границ и середин интервалов

Первый интервал будет лежать в пределах $y_{\min} \dots y_1$, где

$$y_1 = y_{\min} + h. \tag{3}$$

Для каждого интервала определяют его середину y_i^* по формуле

$$y_i^* = \frac{y_{i-1} + y_1}{2}, \tag{4}$$

где y_{i-1} – нижняя граница интервала;
y_i – верхняя граница интервала.

Пример

$y_1 = 16{,}7 + 6{,}29 = 22{,}99$ кг/м^3 и т. д.

Так, для первого интервала середина будет равна:

$$y_1^* = \frac{16{,}7 + 22{,}99}{2} = 19{,}85 \text{ кг/м}^3.$$

Для систематизации расчетов их можно свести в таблицу.

Сводная таблица расчетов

№ интер-вала	Границы интервала	Середина интер-вала y_i^*	Число значений y_i, попадающих в интервал m_i	Относи-тельная частота p_i	$p_i/h \cdot 10^{-3}$
1	16,7…22,99	19,85	7	0,12	19,08
2	22,99…29,28	26,14	6	0,10	15,90
3	29,28…35,57	32,43	9	0,15	23,85
4	35,57…41,86	38,72	4	0,07	11,13
5	41,86…48,15	45,01	4	0,07	11,13
6	48,15…54,44	51,30	16	0,27	42,93
7	54,44…60,7	57,57	14	0,22	34,98

4.4. Подсчет числа результатов эксперимента, попадающих в каждый интервал

Данный подсчет можно проводить визуально.

Предварительно необходимо принять решение о том, к какому интервалу всегда будут относить значение случайной величины, попадающее на границу интервала.

Сумма всех результатов эксперимента, попадающих в каждый интервал, должна быть равна объему выборки.

$$n = \sum_{i=1}^{k} m_i , \tag{5}$$

где m_i – число результатов эксперимента, попадающих в i-й интервал.

Пример

7 + 6 + 9 + 4 + 4 + 16 + 14 = 60 (см. таблицу).

4.5. Расчет относительной частоты попадания случайной величины в каждый интервал

Относительная частота (вероятность) попадания случайной величины в каждый интервал рассчитывается по формуле

$$p_i = \frac{m_i}{n} , \tag{6}$$

$$\sum_{i=1}^{k} p_i^* = 1 . \tag{7}$$

Пример

Так, для первого интервала относительная частота равна:

$$p_i = \frac{7}{60} = 0{,}12\,.$$

4.6. Расчет основных статистических показателей выборки

К основным статистическим показателям относятся объем выборки n, выборочное среднее $\bar{y}$ и выборочная дисперсия S^2, которые определяются по формулам

$$\bar{y} = \frac{\sum_{i=1}^{k} y_i m_i}{n}, \tag{8}$$

$$S^2 = \frac{\sum_{i=1}^{k} m_i \left(y_i^* - \bar{y} \right)^2}{n-1}\,. \tag{9}$$

Пример

Так, для приведенной выше выборки, используя данные таблицы, можно рассчитать

$$\bar{y} = \frac{19{,}85 \cdot 7 + 26{,}14 \cdot 6 + 32{,}43 \cdot 9 + 38{,}72 \cdot 4 + 45{,}01 \cdot 4 + 51{,}30 \cdot 16 + 57{,}59 \cdot 14}{60} = 42{,}49 \text{ кг/м}^3.$$

$$S^2 = \frac{7(19{,}85 - 42{,}49)^2 + 6(26{,}14 - 42{,}49)^2 + 9(32{,}43 - 42{,}49)^2 + 4(38{,}72 - 42{,}49)^2 +}{60-1}$$

$$\frac{+4(45{,}01 - 42{,}49)^2 + 16(51{,}30 - 42{,}49)^2 + 14(57{,}59 - 42{,}49)^2}{60-1} = 2030{,}42 \text{ кг/м}^3.$$

4.7. Графическое представление статистического ряда

График, построенный по данным статистического ряда, называется **гистограммой**. Для его построения по оси абсцисс откладывают значения границ интервалов и на каждом из них, как на основании, строят прямоугольник, площадь которого должна быть равна относительной частоте, соответствующей данному интервалу. Следовательно, высота каждого прямоугольника равна p_i / h (рис. 1).

Кривая распределения случайной величины представляет собой зависимость между ее значениями и соответствующими им вероятностями. В данном случае она строится по серединам интервалов y^* и соответствующей относительной частоте p_i (рис. 2).

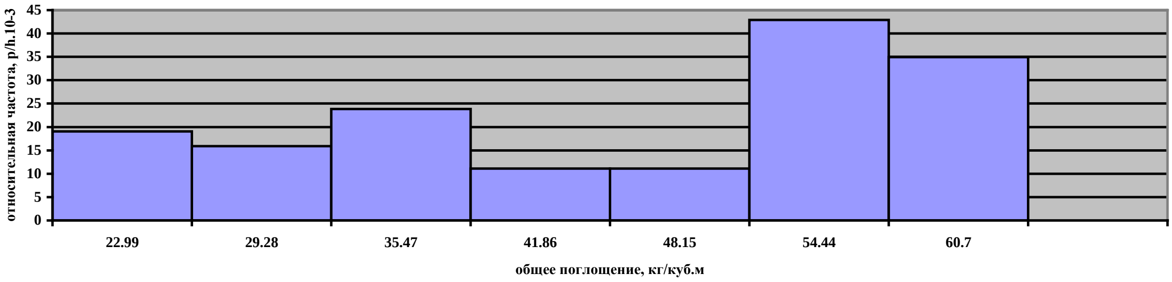

Рис. 1. Гистограмма

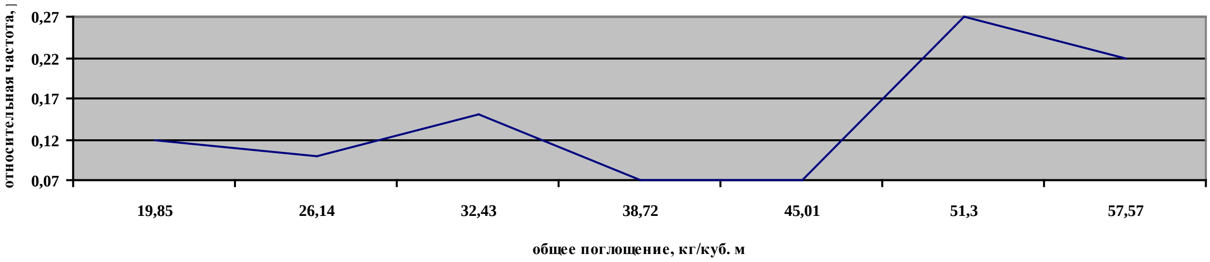

Рис. 2. Кривая распределения общего поглощения

4.8. Выводы

В выводах необходимо перечислить основные статистические показатели, характеризующие исследуемую выборку, а также указать, подчиняется ли распределение случайных величин в выборке закону нормального распределения или нет.

Пример

Исследуемую выборку характеризуют следующие основные статистические показатели:

– выборочное среднее $\bar{y} = 42{,}49$ кг/м^3;

– выборочная дисперсия $S^2 = 2030{,}42$ кг/м^3;

– объем выборки $n = 60$.

Распределение случайных величин в выборке не подчиняется закону нормального распределения (см. рис. 1, 2).

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Пижурин, А. А. Основы научных исследований в деревообработке : учебник для студентов вузов, обучающихся по дневной и заочной форме специальностей 250403 (260200) «Технология деревообработки» и 150405 (170400) «Машины и оборудование лесного комплекса» / А. А. Пижурин, А. А. Пижурин ; Моск. гос. ун-т леса. – М. : МГУЛ, 2005. – 305 с.

2. Пижурин, А. А. Научные исследования в деревообработке : Основы научных исследований : текст лекций : для студентов специальностей 2602.00 и 1704.00 / А. А. Пижурин ; [ред. Е. Г. Петрова] ; Моск. гос. ун-т леса. – М. : МГУЛ, 1999. – 103 с.